黄建峰 2120150994

# 3、数据分析过程及结果

## 3.1、数据摘要

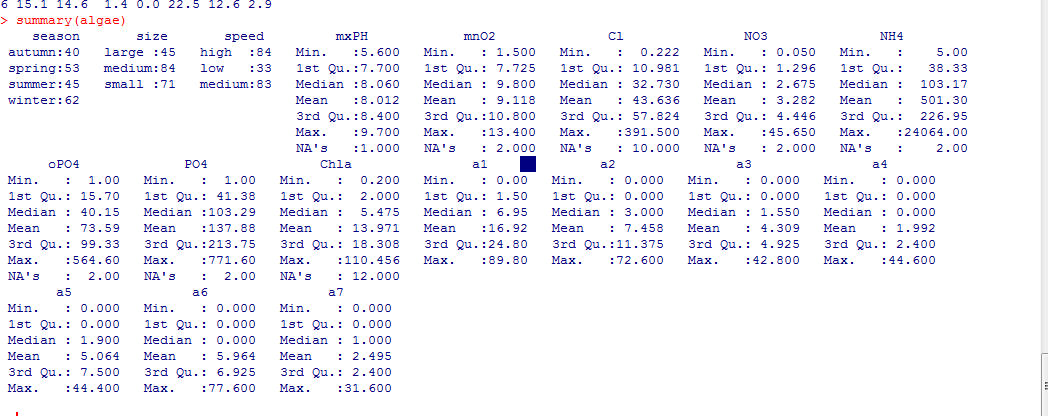
### （1）加载数据

利用R软件载入DMwR添加包，里面有我们需要的名为algae的数据框。这个数据框里含有前面提到的200个观测值：

> library(DMwR)

### （2）显示数据摘要信息

控制台下输入：> summary(algae)



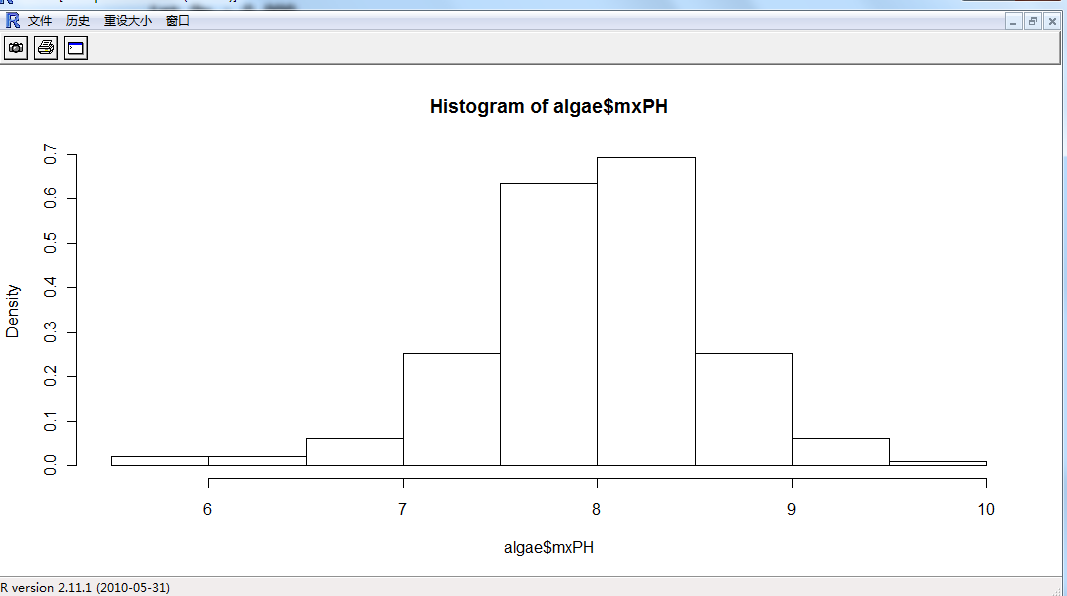
这个简单函数立即给出了数据的的统计特征概括，对于标称变量，它给出了每个取值的变量的频数。例如，从结果中可知冬季采集的水样比其他季节更多，河流size为最大的有45个，河流的流速low比较少。对于数值型变量，给出了四分位数，中位数，均值，极值，缺失数等信息。这些统计信息提供了变量值分布的初步信息，在变量有缺失值的情况下，字符串NA后面的数值即为缺失值的个数，通过中位数，均值，四分位数的信息，我们可以了解数据分布的偏度和分散情况。

## 3.2、数据可视化

### （1）绘制直方图

输入：> hist(algae$mxPH,probability = T)

该程序绘制变量mxPH的直方图。如下图，变量mxPH的分布非常接近正态分布，它的值大多数都集中在变量的均值附近。



### （2）绘制qq图

用Q-Q图来检验mxPH是否是正态分布，加载包car，使用里面的qq.plot()绘制Q-Q图，代码如下：

> library(car)

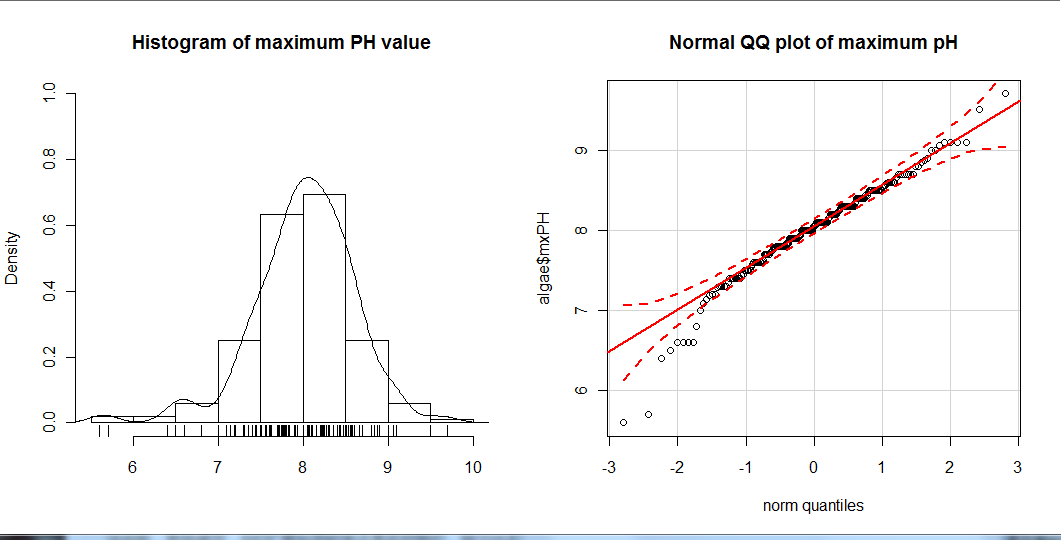
> par(mfrow=c(1,2))

> hist(algae$mxPH,prob=T,xlab = "",ylim = 0:1,main = "Histogram of maximum pH value")

> lines(density(algae$mxPH,na.rm=T))

> rug(jitter(algae$mxPH))

> qq.plot(algae$mxPH,main="Normal QQ plot of maximum pH")



上图的左边是添加核密度曲线的直方图，而在X轴附近绘制变量的实际值，从而容易辨别离群点，例如我们可以发现有两个字显著低于其它所有值。这种数据检查是很重要的，因为它能确定数据样本中可能出现的错误。右边的Q-Q图，他绘制了变量值与正态分布的理论分位数的散点图，同时他给出正态分布的95%的置信区间的带状图，从右上图知，变量有几个小的值明显在95%置信区间之外，它们不服从正态分布。

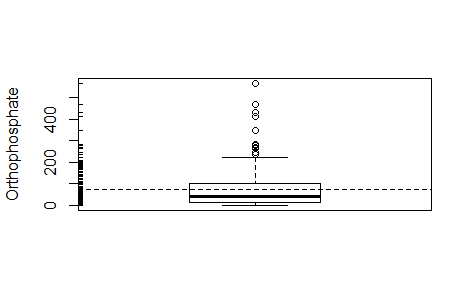
### （3）绘制盒图

用盒图来检验变量OP04，r代码如下

> boxplot(algae$oPO4,ylab="Orthophosphate")

> rug(jitter(algae$oPO4),side = 2)

> abline(h=mean(algae$oPO4,na.rm = T),lty=2)



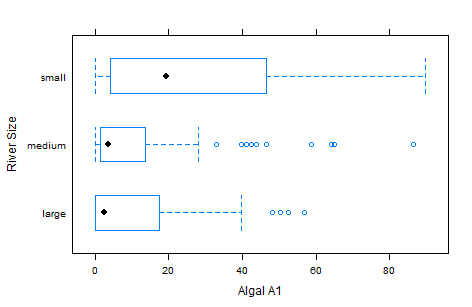
盒图上方小横线上面的小圆圈表示与其他值比较特别大的值，通常认为是离群值，这意味着盒线图给出大量的信息，它不仅给出变量的中心趋势，也给出了变量的发散情况和离群值。

上图中与X轴平行的直线，是变量的均值所在的位置，将均值线与中位数线进行比较，就可以知道变量的多个离群值使得作为变量中心的均值产生了扭曲。

### （4）绘制条件盒图

采用条件盒图来研究海藻变量的值得分布，r代码如下：

bwplot(size~a1,data=algae,ylab = "River Size",xlab = "Algal A1")



从图知，在规模较小的河流中，海藻的频率越高。

# 4、数据缺失处理

## 4.1 将缺失部分剔除

剔除含有缺失数据的记录非常容易实现，尤其是当这些记录所占的比例在可用数据集中非常小的时候，这个选择就比较合理。

为了从数据框中剔除16个样本，我们可以简单的输入：

algae <- na.omit(algae)

## 4.2用高频率值来填补缺失值

填补缺失数据最简单和快捷的方法是使用一些代表中心趋势的值，代表中心趋势的值反映了变量分布的最常见的值，因此中心趋势值是最自然的选择。有多个代表数据中心趋势的指标，例如平均值，中位数，众数等。最合适的选择由变量的分布决定。对于接近正态的分布来说，所有的观测值都较好地聚集在平均值周围，平均值数就是最佳选择。然而，对于偏态分布，或者离群值的变量来说，选择平均值就不好。偏态分布的大部分值都聚集在变量分布的一侧，因此平均值不能作为最常见的代表。另一方面，离群值（极值）的存在会扭曲平均值，这就导致了平均值不具有代表性的问题。因此，在对变量分布进行检查之前选择平均值作为中心趋势的代表是不明智的，例如某些R的绘制工具。对偏态分布或者有离群值的分布而言，中位数是更好的代表数据中心趋势的指标。

代码如下：

algae[48,'mxPH'] <- mean(algae$mxPH,na.rm=T)

algae[is.na(algae$Chla),'Chla'] <- median(algae$Chla,na.rm=T)

## 4.3通过变量的相关关系来调补缺失值

另一种获取缺失值较少偏差估计值的方法是探寻变量之间的相关关系。比如，通过变量值之间的相关关系，能够发现某变量与mxPH高度相关。这可以使我们得到含有缺失值的第48条样本的更可能的填补值。这比之前使用平均值的方法将更胜一筹。

> options(digits = 1)

> cor(algae[,4:18],use="complete.obs")

函数cor（）的功能是产生变量之间的相关值矩阵（因为前3个变量时名义变量，所以计算相关值时不考虑它们）。设定参数use=“complete.obs”时，R在计算相关值时忽略含有NA的记录。相关值在1（或-1）周围表示相应的两个变量之间有强正（或负）线性相关关系。然后其他R函数可以得到变量间线性相关的近似函数形式，它可以让我们通过一个变量的值计算出另一个变量的值。

## 4.4、通过探索案例之间的相似性来填补缺失值

不同于探索数据集列之间的相关性，尝试使用行之间的相似性来填充缺失值。我们可以使用这种方法来填补剔除了那两个含有太多的NA值得样本外的其它缺失数据。

如上所描述的方法假如两个水样是相似的，其中某些变量有缺失值，那么该缺失值很可能与另一个水样的值是相似的。为了使用这种直观的方法，首先定义相似性的概念。相似性经常由描述观察值的多元度量空间的变量所定义。在文献中有许多度量相似性的指标，常用的是欧式距离。这个距离可以非正式的定义为任何两个案例的观察值之差的平方和，计算公式如下：

下面要描述的方法是使用这种度量来寻找与任何含有缺失值的案例最相似的10个水样，并用它们来填充缺失值。我们考虑两种应用这些值的方法。第一种方法简单地计算这10个最相近的案例的中位数并用这个中位数来填充这些缺失值。第二种方法是采用这些相似数据的加权均值。权重的大小随着距待填补缺失值的个案的距离为d，则它的值在加权平均中的权重为：

上面的方法可以通过添加包函数knnImputation()来实现，这个函数用一个欧式距离的变种来找到距离任何个案最近的k个邻居。在计算距离时都要对数据进行标准化。

下面介绍如何使用knnImputation()函数

> algae<-algae[-manyNAs(algae),]

> algae<-knnImputation(algae,k=10)

> algae

如果使用中位数来填补缺失值，可以使用下面的代码

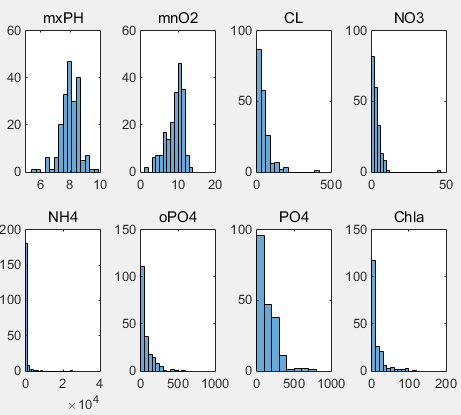
> algae<-knnImputation(algae,k=10,meth = "median")

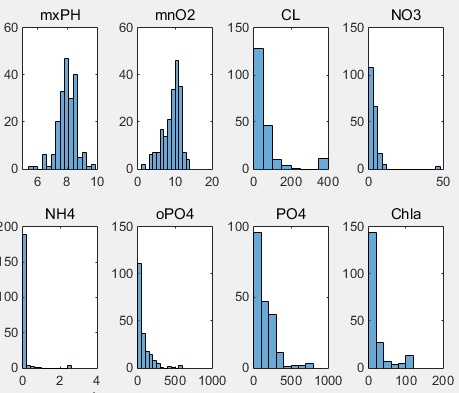
总之使用这些简单地操作，数据中不再含有缺失值NA，未使用R的其他函数进行分析做好充分的准备。

当决定使用哪种方法来填充缺失值时，大多数时候根据你所分析的领域知识来确定，根据个案之间的相似性来填补缺失看起来合理，但这种方法也存在其他问题，例如可能存在不相关的变量扭曲相似性，甚至造成大型数据集的计算特别复杂的问题。另外，对于这些大数据集问题，可以通过随机抽取样本的方法来计算他们之间的相关性。

## 4.5结果对比

预处理后的数据集也可以进行可视化，例如将极值填补后的结果与原直方图相比：





上图为处理前，下图处理后，能看出NO3，NH4，Chla有明显的变化。

其余可视化属性与此同理。